Proceedings of the Canadian Society for Mechanical Engineering International Congress 2023
CSME Congress 2023
May 28-31, 2023, Sherbrooke, QC, Canada

Developing Machine Learning Models to Predict Methane and Nitrogen Oxide Engine-out Emissions from a Heavy-duty Natural-gas Engine

Navid Balazadeh Meresht[1*], Sandeep Munshi[2], Mahdi Shahbakhti[3], Gordon McTaggart-Cowan[1]
[1]School of Sustainable Energy Engineering, Simon Fraser University, Surrey, Canada
[2]Westport Fuel Systems Inc, Vancouver, Canada
[3] Mechanical Engineering Department, University of Alberta, Edmonton, Canada
*navid_balazadeh_meresht@sfu.ca

Abstract—Heavy-duty engine manufacturers must comply with challenging and more stringent emission and greenhouse gas (GHG) regulations. Predicting engine emission behavior in system-level models with reasonable accuracy is advantageous for engine and powertrain development. Machine learning (ML) models are promising alongside 3D physics-based and one-dimensional models. In this study, five different ML models are trained using experimental engine data for emission prediction of methane (CH_4) and nitrogen oxides (NOx). The models are compared with an existing phenomenological engine model (GT-Power). The ML models include linear regression, Ridge regression, Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). The results show that the RF model outperforms other models and a one-dimensional model regarding NOx emission prediction. The results of RF NOx and CH_4 emission prediction in the test set fit with 80% accuracy (±20 error margin). Also, 95% of test data points have less than 10% error compared to real experimental data.

Keywords-Natural gas engine; Methane emissions; Nitrogen oxide emissions, one-dimensional simulation, Machine learning, Ridge, SVM, Random Forest, XGBoost

I. INTRODUCTION

Road freight transport (long-hauling) is the backbone of European trade, generating more than €334 billion turnover and providing jobs for 3.2 million Europeans [1]. In North America, trucks are responsible for 66% of the trade between the US, Canada, and Mexico. Tractor-trailers (class 8 commercial trucks) emit a significant part of greenhouse gases (GHGs) and nitrogen oxide (NOx) emissions, although this class forms a small part of commercial trucks. For example, class-8 trucks form 9% of the commercial vehicles fleet in the US but emit almost half of the NOx and GHGs [2]. Decarbonization in commercial vehicles and heavy-duty fleets is generally driven by stringent carbon dioxide (CO_2) regulations. In this regard, different studies focus on efficiency improvements of commercial vehicles in three parts: engine, vehicle, and powertrain. Also, more reduction can be achieved if the technology is combined with low-carbon fuels such as natural gas. In fact, alternative fuels, such as natural gas, can provide a short-term solution to reduce GHGs. For example, pilot-ignited direct injection of natural gas engines, which use a small amount of diesel for ignition, offer diesel-like efficiency and performance but 15-20% lower GHGs. Evaluation tools play an important role in predicting the emission behavior associated with different technologies. Predicting emissions with reasonable accuracy leads to assessing strategies for engine emission enhancement, after-treatment development, and compliance with the most recent emission regulations. This study aims to utilize different ML regressors as predictive models to estimate the methane (CH_4) and NOx emissions of a natural gas heavy-duty engine. Moreover, this study compares the performance of those models and an existing one-dimensional engine model. Finally, one of the future aims of this study is to use the one-dimensional model to provide the ML models with the required inputs (features) since the one-dimensional model performs well in predicting those features. As a result, several test cycles can be examined with the ML and phenomenological model without the need for experimental tests.

Physics-based models have been widely used to model engine combustion and emission behavior [3], [4]. Detailed 3D simulation models using physical insight can reproduce physical phenomena and model the engine. However, the main drawback of 3D CFD is the high computational cost, which makes them impractical in model-based control and calibration systems. Also, 1D simulation tools can reasonably predict engine performance, but they are less accurate in complex emission modelling. NOx emissions are very sensitive to the maximum cylinder temperature. However, the 1D tool considers two temperature zone for the calibration process. In particular, as they do not have a spatial resolution of temperature or concentration fields, accurate prediction of partial combustion byproduct emissions and unburned fuel can be challenging, especially in non-premixed combustion. On the other hand, predicting engine emissions with ML regression models with respect to engine operating conditions (different features) would

be very helpful, especially once they are combined with simulation models. The benefits include lowering computation costs and increasing the applicability of the models in real-time control systems.

Recently, ML models and their diverse applications in engine development and calibration/control have received much interest. A comprehensive review showed that when applied appropriately, ML could be a powerful tool in engine modelling, diagnostics, control, and optimization [5]. The authors suggested grey-box models as a good solution that combines the benefits of ML-based and physics-based models simultaneously. Some studies also focused on ML applications to predict emissions in real-world conditions for diesel vehicles. Cornec et al. [6] used a clustering technique to categorize 70 diesel vehicles based on their emission behavior. Then, they applied non-linear regression and a neural network multi-layer perceptron (MLP) to predict NOx emissions; relative errors in both models are less than 20%. Jiaqiang et al. [7] utilized a deep learning differentiation model after reducing the noise in data using Singular Spectrum Analysis (SSA). They used GRU (Gated Recurrent Unit) to complete high-frequency sequences and the Support Vector Regressor (SVR) for low—frequency segments. The results showed that the combined approach could predict the NOx emissions better than a single model. Another study showed the ability of SVR and Gaussian process regression (GPR) 's ability to predict NOx, CO_2, and fuel consumption in real-world conditions [8]. SVM capability in emission prediction is also studied in similar papers; Norouzi et al. inspected the SVM method to predict engine-out NOx emissions in a diesel engine. SVM contributed to emission prediction, which was finally embedded in a control-oriented model for NOx control and reduction [9]. SVM benefit in NOx prediction is also shown in another similar paper by Ramezani et al. [10].

Although some studies focus on implementing ML models for emission prediction in diesel engines, such work is rare in heavy-duty natural gas (NG) engines. As a result, this study tries to address this knowledge gap by testing different ML models for NOx and CH_4 emission prediction in a heavy-duty natural gas engine. For NOx emissions, a separate phenomenological one-dimensional engine model had been previously developed in GT-Power. The NOx prediction capability of the ML model is compared to that of the calibrated phenomenological combustion model. The NOx prediction in the 1-D model is limited by the two-zone combustion model, which is less accurate over some operating conditions. The ML model has the potential to increase the NOx prediction accuracy under steady and transient operating conditions.

II. METHODOLOGY

A. Experimental setup

In this study, a 13 L NG heavy-duty engine is used for experimental testing and collecting the engine-out CH_4 and NOx emissions. The detailed specification of the engine is written in Table I. The engine is diesel-pilot natural gas, which means only a tiny amount of diesel is injected to prepare the chamber for the main natural gas fuel injection. The experimental engine data was collected by the industrial partner on a multi-cylinder research engine supporting their technology development activities.

TABLE I. Engine Specifications

Engine configuration	Inline six
Advertised power, HP	455
Peak torque, N.m@ rpm	2400@ 1050
Injection system	Pilot diesel, main natural gas injection
Aspiration	Turbocharged
Displacement, L	12.8
Compression ratio	17:1
Bore × stroke, mm × mm	131 × 158
After-treatment systems	DOC/ DPF/ SCR/ ASC

Data was collected at a 0.1 s frequency and included intake and exhaust pressure, temperature and flow measurements, and accurate measurement of fuel flow (NG and diesel). Emissions were measured using an AVL emissions bench measuring undiluted engine-out emissions (i.e., before the after-treatment system).

B. Test cycles

As shown in Fig. 1, world harmonized transient cycle (WHTC) is more transient with more fluctuations, while world harmonized stationary cycle (WHSC) has some steady modes and 20-sec ramps. Combining these two cycles ensures that the collected data represents the engine's real operation since the data obtained from these two tests covers the main engine operation map. In each second of the cycle, the speed is set in the engine dyno, while air and fuel rate are adjusted to meet the load (torque) demand. In the meanwhile, emissions are measured using a gas analyzer. The main collected engine parameters are engine speed, engine torque, diesel fuel flow, natural gas fuel flow, air flow rate, and exhaust gas temperature after the turbocharger; these six parameters will form the ML input features.

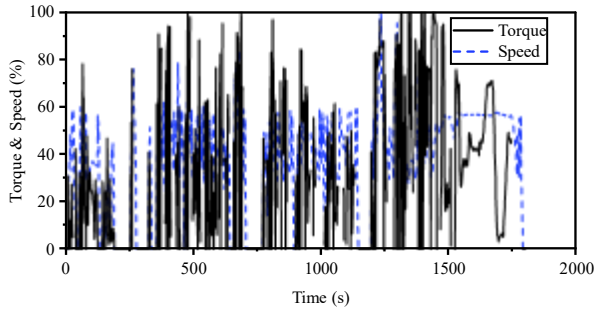
III. GRAY-BOX MODEL

As mentioned in the previous section, six features are selected as the ML model inputs. There were two main reasons for feature selection; 1- physical understanding of the model because it is known that these features contribute more to NOx and CH_4 formation. 2. The other important reason for selecting these features is that the one-dimensional model performs well in predicting these features (95% accuracy), and one of the main future aims of this study is to feed the ML models with the one-dimensional model, instead of experimental tests. The GT-power model is built on experimental data using genetic algorithm (GA) optimization (the details of the steps for building that model are out of this paper's scope). However, the long-term idea is to use the engine plant model built in GT-Power to provide input parameter values for the ML model.

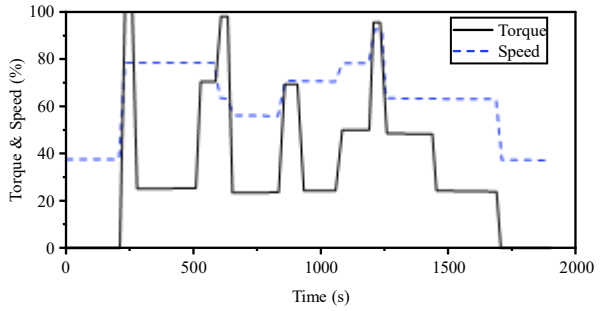
IV. MACHINE LEARNING WORKFLOW

Fig. 2 shows the flowchart for training different machine learning models. After collecting the data, cleaning, scaling,

and stratified sampling, different models are selected to be trained.



(a)



(b)

Figure 1. Engine cycles tested over engine dyno; a) WHTC, b) WHSC

A. *Data cleaning and scaling*

After completing the tests, the data is collected at a 10 Hz frequency. The data then is down-sampled to 1Hz frequency. Data with negative torque values are removed from the collected data since these points do not significantly contribute to the total emissions. In this way, 23% of the original data is excluded from the study. The remaining data are divided into training and testing sets with a random sampling method. Before feeding the data to the ML models, the data must be scaled since the range of different features is completely different. Scaling in this study is performed using standardization. Standardization does not limit the data within a specific range and is much less affected by the outliers.

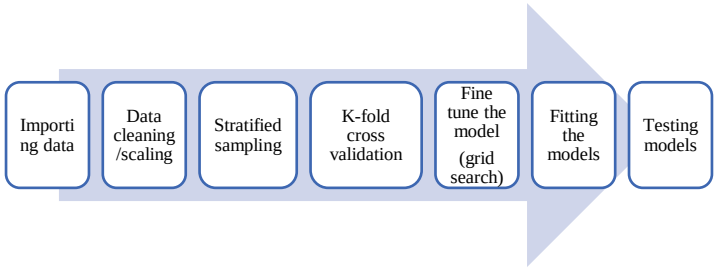


Figure 2. Different steps in this study for training the models

B. *Stratified sampling*

After cleaning the data, the stratified sampling method splits the data to train and test sets. 10% of the cleaned data is allocated to the test set (275 data points), while 90% is to the train (and validation) set (2484 data points). It should be noted that stratified sampling is done with regard to the most significant feature of each label. For example, as shown in Table II for CH_4, this emission correlates well with engine speed. As a result, a stratified sampling process of features has been performed by categorizing engine speed in different bins.

C. *K-fold cross-validation*

K-fold cross-validation is implemented for the models to initialize them and to have an overview of the different models' performance. 20-fold cross-validation is performed in this study for all the involved models. In each iteration, the K-fold algorithm chooses one group as a fold, trains a model on the rest of the groups (out of the fold), and assesses it on the fold set [12].

D. *Grid search*

Hyperparameters such as tolerated error, regularization parameters, and iteration stop criteria significantly impact the model performance. The implemented models and their hyperparameters are introduced in the next section. The grid search technique tries all the possible hyperparameter combinations within a given range to find the best hyperparameters, leading to the minimum error [11]. In this study, 30-fold cross-validation is performed during the grid search algorithm which creates reasonable combinations of hyperparameters and training set.

Table II. Correlation matrix for CH_4 regarding different features

	Speed	Torque	Diesel fuel rate	NG fuel rate	Air flow rate	Post turbine temperature	CH_4
CH_4	0.69	0.12	0.37	0.23	0.45	0.14	1

V. MACHINE LEARNING METHODS

This section introduces the different machine learning methods, their loss (utility) functions, and the associated hyperparameters.

A. *Linear Regression*

Linear regression model prediction is shown in (1) (and in the vectorized form in (2)).

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \quad (1)$$

$$\hat{y} = h_\theta(x) = \theta \cdot X \quad (2)$$

Where $\hat{y}$ is the predicted value, n is the number of features, x_i is the i^{th} feature value, and θ_j is the j^{th} model parameter. Also, the mean squared error cost function for a linear regression model is shown in (3) [13].

$$MSE\ (X, h_\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\theta^T X^{(i)} - y^{(i)}\right)^2 \quad (3)$$

B. *Ridge Regression*

Ridge regression is one of the regularized versions of linear regressions which adds $\alpha \sum_{i=1}^{n} \theta_i^{\,2}$ to the cost function.

Hyperparameter α controls the amount of regularization. (4) shows the cost function of the Ridge regression model [13].

$$J(\theta) = MSE(\theta) + \frac{\alpha}{2}\sum_{i=1}^{n}{\theta_i}^2 \quad (4)$$

C. SVM

SVM tries to find a correlation between input-output by solving the cost function in (5).

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}{\theta_i}^2 + C\sum_{i=1}^{n}(\zeta_i^+ + \zeta_i^-) \quad (5)$$

Where ζ_i^+, ζ_i^- are slack variables and help to check the possible infeasibilities of the optimization problem. C is also the regularization parameter. In SVM, instead of dealing with $\hat{y} = h_\theta(x_i)$, kernel function can be replaced, $\hat{y} = h_\theta(\Gamma(x_i))$. This method is called the SVM kernel trick and uses higher dimensions of feature x_i in $\hat{y}$. Different kernel functions exist, such as linear, polynomial, and Gaussian RBF [13], [14]. This study uses the polynomial kernel function.

D. Random Forest

Random Forest (RF) is an ensemble of decision trees and is generally trained with the bagging method, which means that random sampling is performed with replacement. RF includes all the hyperparameters of the decision tree plus those of the bagging classifier. RF regressor is based on a Regression Tree (RT), an iterative process that splits the data into different branches using the Classification and Regression Trees (CART) algorithm. The cost function of CART is as follows:

$$J(\theta) = \frac{n_{left}}{n}MSE_{left} + \frac{n_{right}}{n}MSE_{right} \quad (6)$$

Where,

$$MSE(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$

Also, n_{left} and n_{right} are the left and right branches of the tree. The model's regularization is performed by specifying the minimum number of samples at the leaf node. The maximum depth of the tree is another hyperparameter [13].

E. XGBoost

XGBoost, or extreme gradient boosting, belongs to a family of boosting algorithms and uses the gradient boosting (GBM) framework based on a paper by Friedman [15]. It uses an optimized gradient-boosting algorithm through parallel processing, tree-pruning, handling missing values, and regularization to tackle overfitting. It also has several excellent capabilities, such as handling early stopping. Tianqi Chen initially introduced this package as part of the Distributed (Deep) Machine Learning Community (DMLC) [13], [16].

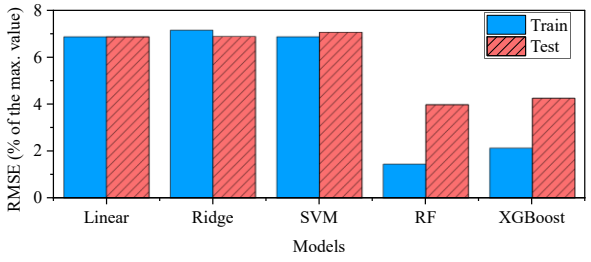
Table III. summarizes the models, their hyperparameters, the range used in grid search to fine-tune the models, and the best-obtained hyperparameters. After finding the best hyperparameter for each of the models using the grid search method (30-fold cross-validation), the best hyperparameter is assigned to the model. Then the model is fitted over the training dataset and predicted the results for the test set.

Table III. Utilized models, their hyperparameters, and the applied ranges in the grid search technique

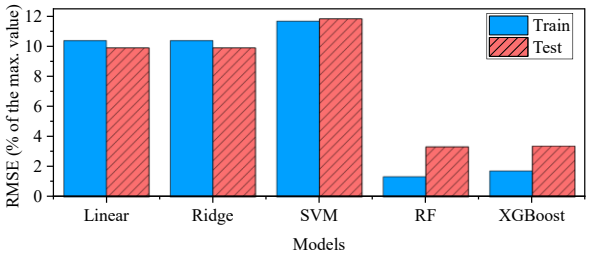
Model	**Hyperparameters**	**Range used in grid search**	**Best hyperparameter**
Linear regression	-	-	-
Ridge regression	α	0 to 5.5 (with 0.5 step)	4.5
SVM	Kernel function Degree C ϵ	Kernel= poly Degree= 1,2,3 C=0.01,0.1,1,10,100 ϵ=0.1, 0.5, 1	Kernel= poly Degree= 1 C=1 ϵ=1
RF	$n_{estimator}$ max_features	$n_{estimator}$=3,10,30,100 max_features=1,2,3,4	$n_{estimator}$=100 max_features=4
XGBoost	$n_{estimator}$ max_depth	$n_{estimator}$=3,10,30,100,200 max_depth=3,5,10	$n_{estimator}$=200 max_depth=10

VI. Results

The best model is used for predicting the test data. Fig. 3 shows the root mean squared error (RMSE) of the models for CH_4 and NOx, both in the training set and test set. The RMSE is scaled as the percentage of the maximum value of the label and is obtained from the best models, i.e., the tuned models with the best hyperparameter. Both for CH_4 and NOx, RF shows the minimum RMSE. Also, RMSE is higher for test sets in RF and XGBoost, but the difference is not considered to conclude the overfitting issue. Also, in RF model, RMSE is less than 4% of the maximum value of the targets (emission values).



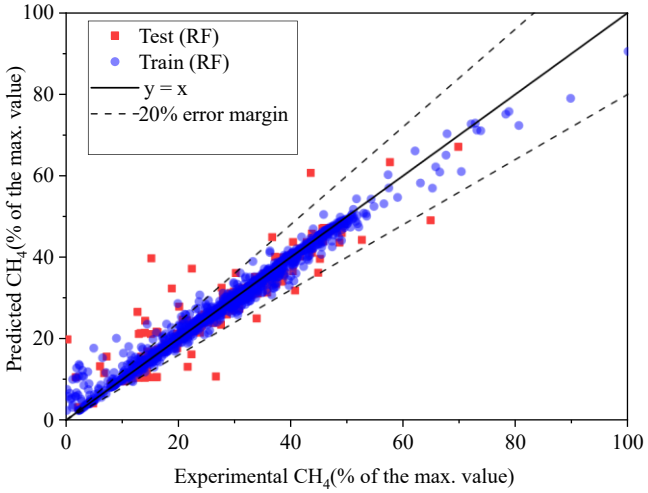
(a)



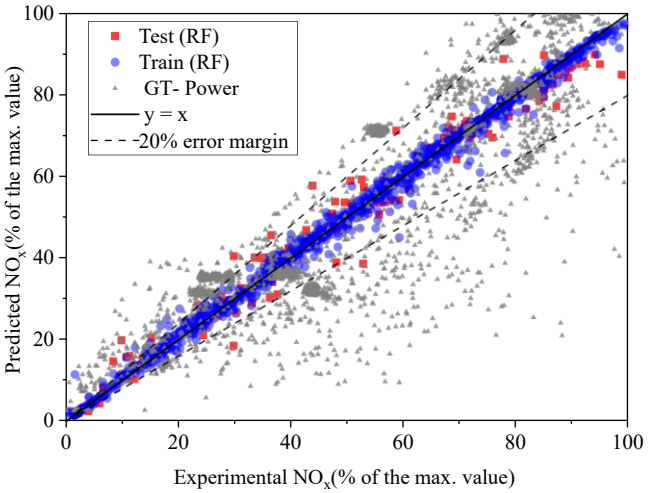
(b)

Figure 3. Scaled RSME (% of the maximum value of the label) for a) CH_4 and b) NOx obtained from the fine-tuned models

The distribution of the CH_4 and NOx data set (both for train and test sets) against the experimental data with RF and SVM is depicted in Fig. 4. As seen in Fig. 4a and Fig. 4b, most data points fit within 80% accuracy with RF (±20% error margin).



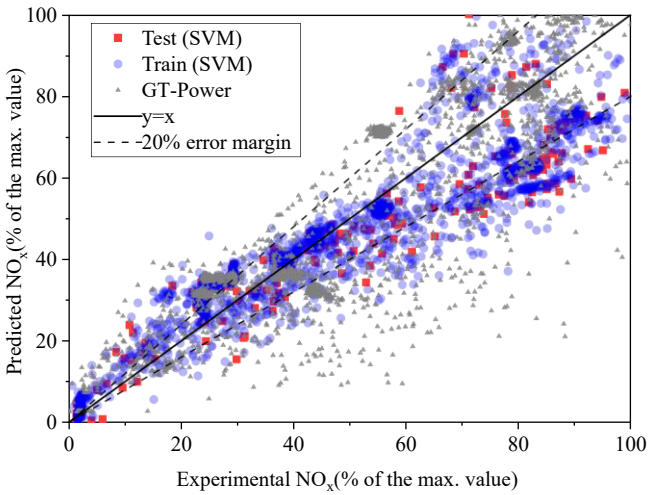
(a)



(b)

Figure 4. Distribution of the data set against the experimental points; a) for CH_4 -RF, b) NOx-RF

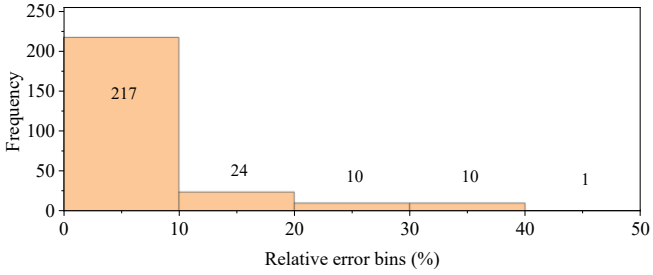
As seen in Fig. 4c, the phenomenological model and SVM have comparable distribution and accuracy, while the RF model has a smaller error margin. Although the prediction of phenomenological model and SVM method overlaps in many data points, SVM shows better accuracy in some parts of the data. In other words, the phenomenological model in those parts underpredict, compared to SVM. Embedding such ML models in the simulation tool can be helpful in reducing the relative error to have more confidence in the prediction.



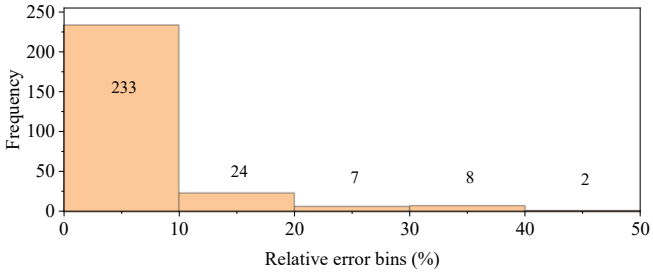
(c)

Figure 4. Distribution of the data set against the experimental points for c) NOx-SVM

Fig. 5 also shows the RF model's error in the test set for CH_4 and NOx prediction. As seen in the error distribution in Fig 5. for RF, 80% to 90% of the test set has less than 10% error. Also, errors higher than 40% might be related to the low torque values that could have been removed in the data cleaning since they can be considered outliers. The distribution of the relative error which is (predicted-test label)/test label, confirms that most predicted points have less than 10% error.



(a)



(b)

Figure 5. Error distribution for the test sets, a) for CH_4 and b) for NOx, obtained from the RF model

VII. Summary and Conclusion

This study uses different machine learning regression models to predict NOx and CH_4 emissions of a heavy-duty natural gas engine. The ML models include linear regression, ridge regression, SVM, RF and XGBoost. The results of different ML models are compared with each other. Also, ML models are compared with the previously developed one-dimensional phenomenological model. The main results are as follows:

- The minimum RMSE values for CH_4 and NOx belong to the RF model, followed by XGboost, Ridge, linear and SVM. Generally, RF yields an overall better model due to its greater tree diversity and the added extra randomness when growing trees.
- SVM generalizes the model slightly better compared to RF. In other words, SVM has lower variance at the cost of added bias since the RMSE of the train and test sets are very close. Although the relative error of SVM is higher than RF, the significance of the added bias (error) in SVM depends on the target application (trade-off between bias and variance). Moreover, the difference in RMSE of train and test sets in RF is not high enough to conclude the overfitting issue.
- ML models show that the accuracy of the prediction is within 80% (±20% relative error margin). Also, when the test data is inspected, most of the points in the test set have less than 10% error.
- Although the hyperparameters are tuned using the grid search method, the assigned range can be finer. Also, more hyperparameters can be added to each model to evaluate other hyperparameters' effect.
- The ML model can be integrated into the 1-D engine simulation to offer higher accuracy prediction of engine-out NOx and CH_4 emissions compared to the phenomenological combustion model's predictions over steady state and transient operation.

Acknowledgment

The authors acknowledge the technical and financial support of Westport™ Fuel Systems and funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Nomenclature

CART	classification and regression trees
GHG	greenhouse gas
GA	genetic algorithm
ML	machine learning
NG	natural gas
RF	random forest
RMSE	root mean squared error
RT	regression tree
SVM	support vector machine
XGBoost	extreme gradient boosting
WHSC	world harmonized stationary cycle
WHTC	world harmonized transient cycle

References

[1] "ACEA - European Automobile Manufacturers' Association." https://www.acea.auto/

[2] A. Joshi, "Characterizing the in-use heavy-duty vehicle fleet," *Mobility Notes*. https://mobilitynotes.com/characterizing-the-in-use-heavy-duty-vehicle-fleet/

[3] B. Singalandapuram Mahadevan, J. H. Johnson, and M. Shahbakhti, "Development of a Kalman filter estimator for simulation and control of particulate matter distribution of a diesel catalyzed particulate filter," *International Journal of Engine Research*, vol. 21, no. 5, pp. 866–884, Jun. 2020, doi: 10.1177/1468087418785855.

[4] Z. Gao and W. Schreiber, "A phenomenologically based computer model to predict soot and NOx emission in a direct injection diesel engine," *International Journal of Engine Research*, vol. 2, no. 3, pp. 177–188, Jun. 2001, doi: 10.1243/1468087011545415.

[5] M. Aliramezani, C. R. Koch, and M. Shahbakhti, "Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques: A review and future directions," *Progress in Energy and Combustion Science*, vol. 88, p. 100967, Jan. 2022, doi: 10.1016/j.pecs.2021.100967.

[6] C. M. A. Le Cornec, N. Molden, M. van Reeuwijk, and M. E. J. Stettler, "Modelling of instantaneous emissions from diesel vehicles with a special focus on NOx: Insights from machine learning techniques," *Science of the Total Environment*, vol. 737, 2020, doi: 10.1016/j.scitotenv.2020.139625.

[7] J. Li, Y. Yu, Y. Wang, L. Zhao, and C. He, "Prediction of transient NOx emission from diesel vehicles based on deep-learning differentiation model with double noise reduction," *Atmosphere*, vol. 12, no. 12, 2021, doi: 10.3390/atmos12121702.

[8] G. M. H. Shahariar *et al.*, "Real-Driving Co2, Nox and Fuel Consumption Prediction Using Machine Learning Approaches," 2022.

[9] A. Norouzi, D. Gordon, M. Aliramezani, and C. R. Koch, "Machine learning-based diesel engine-out nox reduction using a plug-in PD-type iterative learning control," in *4th IEEE Conference on Control Technology and Applications, CCTA 2020, August 24, 2020 - August 26, 2020*, Virtual, Montreal, QC, Canada, 2020, pp. 450–455. doi: 10.1109/CCTA41146.2020.9206277.

[10] M. Aliramezani, A. Norouzi, and C. R. Koch, "Support vector machine for a diesel engine performance and NOx emission control-oriented model," in *21st IFAC World Congress 2020, July 12, 2020 - July 17, 2020*, Berlin, Germany, 2020, vol. 53, no. 2, pp. 13976–13981. doi: 10.1016/j.ifacol.2020.12.916.

[11] A. Mohammad, R. Rezaei, C. Hayduk, T. O. Delebinski, S. Shahpouri, and M. Shahbakhti, "Hybrid Physical and Machine Learning-Oriented Modeling Approach to Predict Emissions in a Diesel Compression Ignition Engine," presented at the SAE WCX Digital Summit, Apr. 2021, pp. 2021-01–0496. doi: 10.4271/2021-01-0496.

[12] "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation." https://ieeexplore.ieee.org/abstract/document/5342427/?casa_token=gWtegVn583QAAAAA:1199WZ4i83PfcQuh8drI4IWrVskh9os00a6GEfw23cMa4j4bWM_5rEUaz_UpIRZ8EMmvvGBPVJA (accessed Dec. 09, 2022).

[13] "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition [Book]." https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ (accessed Dec. 10, 2022).

[14] S. Shahpouri, A. Norouzi, C. Hayduk, R. Rezaei, M. Shahbakhti, and C. R. Koch, "Hybrid Machine Learning Approaches and a Systematic Model Selection Process for Predicting Soot Emissions in Compression Ignition Engines," *Energies*, vol. 14, no. 23, Art. no. 23, Jan. 2021, doi: 10.3390/en14237865.

[15] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[16] "XGBoost + k-fold CV + Feature Importance." https://kaggle.com/code/prashant111/xgboost-k-fold-cv-feature-importance (accessed Dec. 10, 2022).